United Nations A/73/348

General Assembly

Distr.: General
29 August 2018

Original: English

Seventy-third session
Item 74 (b) of the provisional agenda*
Promotion and protection of human rights: human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms

Promotion and protection of the right to freedom of opinion and expression**

Note by the Secretary-General

The Secretary-General has the honour to transmit to the General Assembly the report prepared by the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, submitted in accordance with Human Rights Council resolution 34/18. In the present report, the Special Rapporteur explores the implications of artificial intelligence technologies for human rights in the information environment, focusing in particular on rights to freedom of opinion and expression, privacy and non-discrimination.

* A/73/150.

** The present report was submitted after the deadline in order to reflect the most recent developments.

18-14238 (E) 190918

Please recycle

Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression

Contents

		Page
I.	Introduction	3
II.	Understanding artificial intelligence	3
	A. What is artificial intelligence?	3
	B. Artificial intelligence and the information environment	6
III.	A human rights legal framework for artificial intelligence	10
	A. Scope of human rights obligations in the context of artificial intelligence	10
	B. Right to freedom of opinion	11
	C. Right to freedom of expression	12
	D. Right to privacy	13
	E. Obligation of non-discrimination	14
	F. Right to an effective remedy	15
	G. Legislative, regulatory and policy responses to artificial intelligence	16
IV.	A human rights-based approach to artificial intelligence	17
	A. Substantive standards for artificial intelligence systems	18
	B. Processes for artificial intelligence systems	19
V.	Conclusions and recommendations	21

I. Introduction

1. Artificial intelligence is increasingly influencing the information environment worldwide. It enables companies to curate search results and newsfeeds as well as advertising placement, organizing what users see and when they see it. Artificial intelligence technologies are used by social media companies to help moderate content on their platforms, often acting as the first line of defence against content that may violate their rules. Artificial intelligence recommends people to friend or follow, news articles to read and places to visit or eat, shop or sleep. It offers speed, efficiency and scale, operating to help the largest companies in the information and communications technology sector manage the huge amounts of content uploaded to their platforms every day. Artificial intelligence technologies may enable broader and quicker sharing of information and ideas globally, a tremendous opportunity for freedom of expression and access to information. At the same time, the opacity of artificial intelligence also risks interfering with individual self-determination, or what is referred to in the present report as "individual autonomy and agency".[1] A great global challenge confronts all those who promote human rights and the rule of law: how can States, companies and civil society ensure that artificial intelligence technologies reinforce and respect, rather than undermine and imperil, human rights?

2. The present report does not pretend to be the last word in artificial intelligence and human rights. Rather, it tries to do three things: define key terms essential to a human rights discussion about artificial intelligence; identify the human rights legal framework relevant to artificial intelligence; and present some preliminary recommendations to ensure that, as the technologies comprising artificial intelligence evolve, human rights considerations are baked into that process. The report should be read as a companion to my most recent report to the Human Rights Council (A/HRC/38/35), in which a human rights approach to online content moderation was presented.[2]

II. Understanding artificial intelligence

A. What is artificial intelligence?

3. Artificial intelligence is often used as shorthand for the increasing independence, speed and scale connected to automated, computational decision-making. Artificial intelligence is not one thing only, but rather refers to a "constellation" of processes and technologies enabling computers to complement or replace specific tasks otherwise performed by humans, such as making decisions and solving problems.[3] "Artificial intelligence" can be a problematic term, suggesting as it does that machines can operate according to the same concepts and rules of human intelligence. They cannot. Artificial intelligence generally optimizes the work of

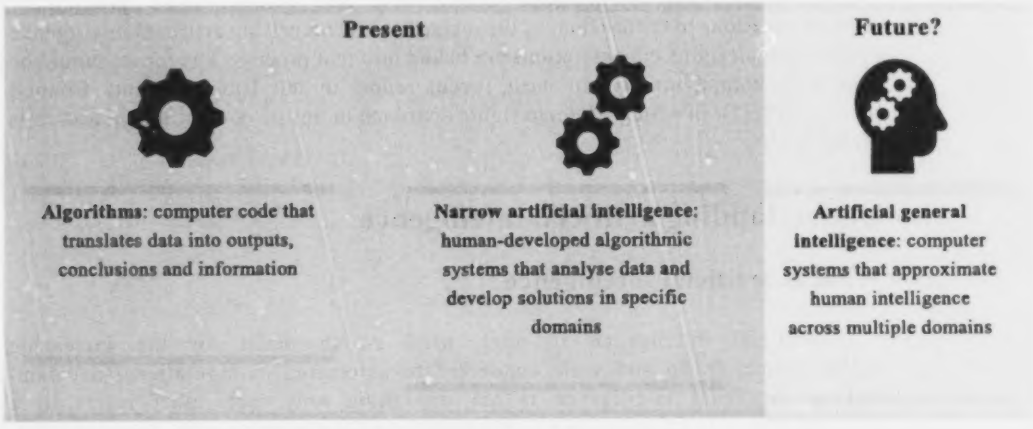
[1] See Mariarosaria Taddeo and Luciano Floridi, "How AI can be a force for good", *Science*, vol. 361, No. 6404 (24 August 2018). Available at http://science.sciencemag.org/content/361/6404/751.full.

[2] The present report benefited from an expert consultation conducted in Geneva in June 2018, supported with a grant from the European Union, and the input from experts as part of the development of document A/HRC/35/38 in 2017 and 2018. I especially wish to thank Carly Nyst and Amos Toh, who contributed essential research and drafting to this project.

[3] See AI Now, "The AI now report: the social and economic implications of artificial intelligence technologies in the near term", 2016. Available at https://ainowinstitute.org/AI_Now_2016_Report.pdf; United Kingdom of Great Britain and Northern Ireland House of Lords Select Committee on Artificial Intelligence, "AI in the United Kingdom: ready, willing and able?", 2018.

computerized tasks assigned by humans through iterative repetition and attempt. That said, it is the language of the culture, of companies and of governments, and I use it here.

4. Popular culture often suggests that society is headed towards artificial general intelligence, a still-distant capability (the "singularity") for a computer system to approximate or surpass human intelligence across multiple domains. [4] For the foreseeable future, there will continue to be advancements with respect to narrow artificial intelligence, according to which computer systems perform programmed tasks (human-developed algorithms) in specific domains. Narrow artificial intelligence underpins, for example, voice assistance on mobile devices and customer service chatbots, online translation tools and self-driving cars, search engine results and mapping services. Machine-learning is a category of narrow artificial intelligence techniques used to train algorithms to use datasets to recognize and help solve problems. For example, artificial intelligence-powered smart home devices are continuously "learning" from data collected about everyday language and speech patterns in order to process and respond to questions from their users more accurately. In all circumstances, humans play a critical role in designing and disseminating artificial intelligence systems, defining the objectives of an artificial intelligence application and, depending on the type of application, selecting and labelling datasets and classifying outputs. Humans always determine the application and use of artificial intelligence outputs, including the extent to which they complement or replace human decision-making.


Present
Future?
Algorithms: computer code that translates data into outputs, conclusions and information
Narrow artificial intelligence: human-developed algorithmic systems that analyse data and develop solutions in specific domains
Artificial general intelligence: computer systems that approximate human intelligence across multiple domains

5. At the foundation of artificial intelligence are algorithms, computer code designed and written by humans, carrying instructions to translate data into conclusions, information or outputs. Algorithms have long been essential to the operation of everyday systems of communication and infrastructure. The enormous volume of data in modern life and the capacity to analyse it fuel artificial intelligence. The private sector certainly sees data that way: the more data available to feed algorithms and the better the quality of that data, the more powerful and precise the algorithms can become. Algorithmic systems can analyse huge volumes of data rapidly, enabling artificial intelligence programmes to perform decision-making functions that were previously the domain of humans acting without computational tools.

[4] Article 19 and Privacy International, "Privacy and freedom of expression in an age of artificial intelligence", London, 2018.

Design stage
Humans procure artificial intelligence systems and define their intentions.
Deployment stage
Humans define the input and design the algorithm.
Humans clean and label the input data.
In some cases, humans define the outputs of artificial intelligence systems.
Artificial intelligence systems define the model used and algorithms learn and adapt independently.
Implementation stage
Humans decide how to apply and use the outputs.

Human agency is integral to artificial intelligence, but the distinctive characteristics of artificial intelligence deserve human rights scrutiny with respect to at least three of its aspects: automation, data analysis and adaptability.[5]

6. **Automation**. Automation removes human intervention from parts of a decision-making process, completing specific tasks with computational tools. This can have positive implications from a human rights perspective if a design limits human bias. For example, an automated border entry system may flag individuals for scrutiny based on objective features such as criminal history or visa status, limiting reliance on subjective (and bias-prone) assessments of physical presentation, ethnicity, age or religion. Automation also enables the processing of vast amounts of data at a speed and scale not achievable by humans, potentially serving public safety, health and national security. However, automated systems rely on datasets that, in their design or implementation, may allow for bias and thus produce discriminatory effects. For instance, the underlying criminal history or visa data suggested above itself may incorporate biases. Excessive reliance on and confidence in automated decisions and a failure to recognize this foundational point may in turn undermine scrutiny of artificial intelligence outcomes and disable individuals from accessing remedies to adverse artificial intelligence-driven decisions. Automation may impede the

[5] Council of Europe, *Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications*, Council of Europe study, No. DGI (2017) 12, 2018. Available at https://www.coe.int/en/web/freedom-expression/-/algorithms-and-human-rights-a-new-study-has-been-published.

transparency and scrutability of a process, preventing even well-meaning authorities from providing an explanation of outcomes.[6]

7. **Data analysis**. Vast datasets support most artificial intelligence applications. Any dataset could form the basis of an artificial intelligence system, from Internet browsing habits to data on traffic flows on highways. Some datasets contain personal data, while many others involve anonymized data. The use by artificial intelligence of such datasets raises serious concerns, including regarding their origins, accuracy and individuals' rights over them; the ability of artificial intelligence systems to de-anonymize anonymized data; and biases that may be ingrained within the datasets or instilled through human training or labelling of the data. Artificial intelligence evaluation of data may identify correlations but not necessarily causation, which may lead to biased and faulty outcomes that are difficult to scrutinize.

8. **Adaptability**. Machine-learning artificial intelligence systems are adaptable, as the algorithms that power them are able to progressively identify new problems and develop new answers. Depending on the level of supervision, systems may identify patterns and develop conclusions unforeseen by the humans who programmed or tasked them. This lack of predictability holds the true promise of artificial intelligence as a transformational technology, but it also illuminates its risks: as humans are progressively excluded from defining the objectives and outputs of an artificial intelligence system, ensuring transparency, accountability and access to effective remedy becomes more challenging, as does foreseeing and mitigating adverse human rights impacts.

B. Artificial intelligence and the information environment

9. Artificial intelligence has particularly important, and sometimes problematic, consequences for the information environment, the complex ecosystem of technologies, platforms and private and public actors that facilitate access to and dissemination of information through digital means. Algorithms and artificial intelligence applications are found in every corner of the Internet, on digital devices and in technical systems, and in search engines, social media platforms, messaging applications and public information mechanisms. In keeping with the focus of my mandate, I note the following three applications of artificial intelligence in the information environment that raise concerns:

10. **Content display and personalization**. Social media and search platforms increasingly dominate how individuals access and share information and ideas and how news is disseminated. Algorithms and artificial intelligence applications determine how widely, when and with which audiences and individuals content is shared. Massive datasets that combine browsing histories, user demographics, semantic and sentiment analyses and numerous other factors feed into increasingly personalized algorithmic models to rank and curate information, that is, to show information to individuals or implicitly exclude it. Paid, sponsored or hashtagged content may be promoted to the exclusion or demotion of other content. Social media newsfeeds display content according to subjective assessments of how interesting or engaging content might be to a user; as a result, individuals may be offered little or no exposure to certain types of critical social or political stories and content posted

[6] Council of Europe, *Algorithms and Human Rights*.

to their platforms.[7] Artificial intelligence shapes the world of information in a way that is opaque to the user and often even to the platform doing the curation.

11. Online search is one of the most pervasive forms of artificial intelligence-powered content display and personalization. Search engines deliver results for queries (and complete or predict queries) using artificial intelligence systems that process extensive data about individual and aggregate users. Because poorly ranked content or content entirely excluded from search results is unlikely to be seen, the artificial intelligence applications for search have enormous influence over the dissemination of knowledge.[8] Content aggregators and news sites[9] similarly choose which information to display to an individual based not on recent or important developments, but on artificial intelligence applications that predict users' interests and news patterns based on extensive datasets. Consequently, artificial intelligence plays a large but usually hidden role in shaping what information individuals consume or even know to consume.

12. Artificial intelligence in the field of content display is driving towards greater personalization of each individual's online experience; in an era of information abundance,[10] personalization promises to order the chaos of the Internet, allowing individuals to find requested information. Benefits may include the ability to access information and services in a greater range of languages[11] or information that is more timely and relevant to one's personal experience or preferences. Artificial intelligence-driven personalization may also minimize exposure to diverse views, interfering with individual agency to seek and share ideas and opinions across ideological, political or societal divisions. Such personalization may reinforce biases and incentivize the promotion and recommendation of inflammatory content or disinformation in order to sustain users' online engagement.[12] To be sure, all sorts of social and cultural settings may limit an individual's exposure to information. But by optimizing for engagement and virality at scale, artificial intelligence-assisted personalization may undermine an individual's choice to find certain kinds of content. This is especially so because algorithms typically will deprioritize content with lower levels of engagement, banishing independent and user-generated content into

[7] World Wide Web Foundation, "The invisible curation of content: Facebook's News Feed and our information diets", 22 April 2018. Available at https://webfoundation.org/research/the-invisible-curation-of-content-facebooks-news-feed-and-our-information-diets/.

[8] Council of Europe, *Algorithms and Human Rights*.

[9] For example, see "How Reuters's revolutionary AI system gathers global news," MIT Technology Review, 27 November 2017. Available at https://www.technologyreview.com/s/609558/how-reuterss-revolutionary-ai-system-gathers-global-news/; Paul Armstrong and Yue Wang, "China's $11 billion news aggregator Jinri Toutiao is no fake," *Forbes*, 26 May 2017. Available at https://www.forbes.com/sites/ywang/2017/05/26/jinri-toutiao-how-chinas-11-billion-news-aggregator-is-no-fake/#1d8b97804d8a.

[10] Carly Nyst and Nick Monaco, *State-Sponsored Trolling: How Governments are Deploying Disinformation as Part of Broader Digital Harassment Campaigns* (Palo Alto, Institute for the Future, 2018).

[11] World Wide Web Foundation, "Artificial intelligence: the road ahead in low- and middle-income countries", Washington, D.C., 2017.

[12] Zeynep Tufekci, "YouTube, the great radicaliser", *New York Times*, 10 March 2018. Available at https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html; James Williams, *Stand Out of our Light: Freedom and Resistance in the Attention Economy* (Cambridge, Cambridge University Press, 2018).

obscurity.[13] Savvy actors can exploit rule-based artificial intelligence systems optimized for engagement to gain higher levels of exposure, and by appropriating popular hashtags or using bots, they can achieve outsized online reach to the detriment of information diversity.

13. **Content moderation and removal**. Artificial intelligence helps social media companies to moderate content in accordance with platform standards and rules, including spam detection, hash-matching technology (using digital fingerprints to identify, for instance, terrorist or child exploitation content), keyword filters, natural language processing (by which the nature of the content is assessed for prohibited words or imagery) and other detection algorithms. Artificial intelligence may be used to subject user accounts to warnings, suspension or deactivation on the basis of violations of terms of service or may be employed to block or filter websites on the basis of prohibited domain data or content. Social media companies use artificial intelligence to filter content across the range of their rules (from nudity to harassment to hate speech and so on), although the extent to which such companies rely on automation without human input on specific cases is not known.[14]

14. Support and pressure for increasing the role of artificial intelligence come from both the private and public sectors. Companies claim that the volume of illegal, inappropriate and harmful content online far exceeds the capabilities of human moderation and argue that artificial intelligence is one tool that can assist in better tackling this challenge. According to some platforms, artificial intelligence is not only more efficient in identifying inappropriate (according to their rules) and illegal content for removal (usually by a human moderator) but also has a higher accuracy rate than human decision-making. States, meanwhile, are pressing for efficient, speedy automated moderation across a range of separate challenges, from child sexual abuse and terrorist content, where artificial intelligence is already extensively deployed, to copyright and the removal of "'extremist" and "hateful" content.[15] The European Commission Recommendation on measures to further improve the effectiveness of the fight against illegal content online of March 2018 calls upon Internet platforms to use automatic filters to detect and remove terrorist content, with

[13] Recently, some tech platforms have indicated their intention to move away from "engagement" — driven personalization to personalization which prioritizes the quality of a user's experience online; see Julia Carrie Wong, "Facebook overhauls News Feed in favour of 'meaningful social interactions'", *The Guardian*, 11 January 2018. Available at https://www.theguardian.com/technology/2018/jan/11/facebook-news-feed-algorithm-overhaul-mark-zuckerberg. However, without thorough transparency, reporting and metrics around how AI systems make and implement such assessments, it is difficult to assess whether this change is having a demonstrable effect on internet users' experience.

[14] An Instagram tool, Deep Text, attempts to judge the "toxicity" of the context, as well as permitting users to customize their own word and emoji filers, and also assesses user relationship in a further attempt to establish context (such as whether a comment is just a joke between friends). Andrew Hutchison, "Instagram's rolling out new tools to remove 'toxic comments'", Social Media Today, 30 June 2017. Available at https://www.socialmediatoday.com/social-networks/instagrams-rolling-out-new-tools-remove-toxic-comments.

[15] The United Kingdom reportedly developed a tool to automatically detect and remove terrorist content at the point of upload. See, for example, Home Office, "New technology revealed to help fight terrorist content online", press release, 13 February 2018. See European Commission, proposal for a directive of the European Parliament and of the Council on Copyright in the Digital Single Market, COM(2016) 593 final, art. 13; Letter from the Special Rapporteur to the President of the European Commission, reference No. OL OTH 41/2018, 13 June 2018. Available at https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-OTH-41-2018.pdf.

human review in some cases suggested as a necessary counterweight to the inevitable errors caused by the automated systems.[16]

15. Efforts to automate content moderation may come at a cost to human rights (see A/HRC/38/35, para. 56). Artificial intelligence-driven content moderation has several limitations, including the challenge of assessing context and taking into account widespread variation of language cues, meaning and linguistic and cultural particularities. Because artificial intelligence applications are often grounded in datasets that incorporate discriminatory assumptions,[17] and under circumstances in which the cost of over-moderation is low, there is a high risk that such systems will default to the removal of online content or suspension of accounts that are not problematic and that content will be removed in accordance with biased or discriminatory concepts. As a result, vulnerable groups are the most likely to be disadvantaged by artificial intelligence content moderation systems. For example, Instagram's DeepText identified "Mexican" as a slur because its datasets were populated with data in which "Mexican" was associated with "illegal", a negatively coded term baked into the algorithm.[18]

16. Artificial intelligence makes it difficult to scrutinize the logic behind content actions. Even when algorithmic content moderation is complemented by human review — an arrangement that large social media platforms argue is increasingly infeasible on the scale at which they operate — a tendency to defer to machine-made decisions (on the assumptions of objectivity noted above) impedes interrogation of content moderation outcomes, especially when the system's technical design occludes that kind of transparency.

17. **Profiling, advertising and targeting**. Advances in artificial intelligence have both benefited from and further incentivized the data-driven business model of the Internet, namely, that individuals pay for free content and services with their personal data. With the vast data resources amassed from years of online monitoring and profiling, companies are able to equip artificial intelligence systems with rich datasets to develop ever more precise prediction and targeting models. Today, advertising by private and public actors can be achieved at an individual level; consumers and voters are the subject of "microtargeting" designed to respond to and exploit individual idiosyncrasies.

18. Artificial intelligence-driven targeting incentivizes the widespread collection and exploitation of personal data and increases the risk of manipulation of individual users through the spread of disinformation. Targeting can perpetuate discrimination, as well as users' exclusion from information or opportunities by, for example, permitting targeted job and housing advertisements that exclude older workers,

[16] Commission recommendation of 1 March 2018 on measures to effectively tackle illegal content online (C(2018) 1177 final). Available at https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online; see also Daphne Keller, "Comment in response to European Commission's March 2018 recommendation on measures to further improve the effectiveness of the fight against illegal content online", Stanford Law School, Center for Internet and Society, 29 March 2018. Available at http://cyberlaw.stanford.edu/publications/comment-response-european-commissions-march-2018-recommendation-measures-further.

[17] See Aylin Caliskan, Joanna Bryson and Arvind Narayanan, "Semantics derived automatically from language corpora contain human-like biases", *Science*, vol. 356, No. 6334 (14 April 2017); Solon Barocas and Andrew Selbst, "Big data's disparate impact", California Law Review, vol. 104, No. 671 (2016).

[18] Nicholas Thompson, "Instagram's Kevin Systrom wants to clean up the &#%@! Internet", Wired, 14 August 2017. Available at https://www.wired.com/2017/08/instagram-kevin-systrom-wants-to-clean-up-the-internet/.

women or ethnic minorities.[19] Rather than individuals being exposed to parity and diversity in political messaging, for example, the deployment of microtargeting through social media platforms is creating a curated worldview inhospitable to pluralistic political discourse.

III. A human rights legal framework for artificial intelligence

A. Scope of human rights obligations in the context of artificial intelligence

19. Artificial intelligence tools, like all technologies, must be designed, developed and deployed so as to be consistent with the obligations of States and the responsibilities of private actors under international human rights law. Human rights law imposes on States both negative obligations to refrain from implementing measures that interfere with the exercise of freedom of opinion and expression and positive obligations to promote rights to freedom of opinion and expression and to protect their exercise.

20. With respect to the private sector, States are bound to guarantee respect for individual rights,[20] especially the rights to freedom of opinion and expression, including by protecting individuals from infringing acts committed by private parties (article 2 (1) of the International Covenant on Civil and Political Rights). States can meet this obligation through legal measures to restrict or influence the development and implementation of artificial intelligence applications, through policies regarding the procurement of artificial intelligence applications from private companies by public sector actors, through self- and co-regulatory schemes and by building the capacity of private sector companies to recognize and prioritize the rights to freedom of opinion and expression in their corporate endeavours.

21. Companies also have responsibilities under human rights law that should guide their construction, adoption and mobilization of artificial intelligence technologies (A/HRC/38/35, para. 10). The Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework provide a "global standard of expected conduct for all businesses wherever they operate" (principle 11), including social media and search companies. To adapt the conclusions from the Guiding Principles to the domain of artificial intelligence (ibid., para. 11), the Guiding Principles require that companies, at a minimum, make high-level policy commitments to respect the human rights of their users in all artificial intelligence applications (principle 16); avoid causing or contributing to adverse human rights impacts through their use of artificial intelligence technology and prevent and mitigate any adverse effects linked to their operations (principle 13); conduct due diligence on artificial intelligence systems to identify and address actual and potential human rights impacts (principles 17–19); engage in prevention and mitigation strategies (principle 24); conduct ongoing review of artificial intelligence-related activities, including through stakeholder and public consultation (principles 20–21),

[19] Julia Angwin, Noam Scheiber and Ariana Tobin, "Dozens of companies are using Facebook to exclude older workers from job ads", ProPublica, 20 December 2017. Available at https://www.propublica.org/article/facebook-ads-age-discrimination-targeting; Julia Angwin, Ariana Tobin and Madeleine Varner, "Facebook (still) letting housing advertisers exclude users by race", ProPublica, 21 November 2017. Available at https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

[20] See principle 3 of the Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework (A/HRC/17/31) and A/HRC/38/35, paras. 6–8.

and provide accessible remedies to remediate adverse human rights impacts from artificial intelligence systems (principles 22, 29 and 31).

B. Right to freedom of opinion

22. The freedom to hold opinions without interference is an absolute right, enshrined in article 19 (1) of the International Covenant on Civil and Political Rights and article 19 of the Universal Declaration of Human Rights. It "permits no exception or restriction," whether "by law or other power".[21] In my 2015 report to the Human Rights Council on encryption and anonymity in digital communications (A/HRC/29/32), I observed that the ways in which information is stored, transmitted and secured in the digital age uniquely affect the exercise of the right to hold opinions. Search queries, browsing activities, email and text communications, and documents and mementos held in the cloud — together, these digital activities and records form the fabric of the opinions users hold (ibid., para. 12). Both State and non-State actors may interfere with these mechanics and processes of forming and holding opinions.

23. An essential element of the right to hold an opinion is the "right to form an opinion and to develop this by way of reasoning".[22] The Human Rights Committee has concluded that this right requires freedom from undue coercion in the development of an individual's beliefs, ideologies, reactions and positions.[23] Accordingly, forced neurological interventions, indoctrination programmes (such as "re-education camps") or threats of violence designed to compel individuals to form particular opinions or change their opinion violate article 19 (1) of the Covenant. The Committee has also found that coercive "inducements of preferential treatment" may rise to a level of persuasion that interferes with the right to form and hold opinions (see CCPR/C/78/D/878/1999).

24. The intersection of technology and content curation raises novel questions about the types of coercion or inducement that may be considered an interference with the right to form an opinion. Content curation has long informed the capacity of the individual to form opinions: for example, media outlets elevate particular stories to the front page with the intention of shaping and influencing individual knowledge about significant news of the day. Commercial advertising has also sought to induce favourable opinions of and cultivate desire for particular products and services.

25. The use of artificial intelligence extends and enhances the tradition of content curation on the Internet, providing more sophisticated and efficient means of personalizing and curating content for the user at a scale beyond the reach of traditional media. The dominance of particular modes of artificial intelligence-assisted curation raises concern about its impact on the capacity of the individual to form and develop opinions. For example, a handful of technology companies lay claim to the vast majority of search queries conducted online. Corporate monopoly of the search market makes it extremely difficult for users to opt out of the algorithmic ranking and curation of search results and may also induce users to believe (as companies intend it) that the results generated are the most relevant or objective information available on a particular subject. The lack of transparency about how search criteria are developed and implemented through the use of artificial

[21] Human Rights Committee, general comment No. 34 (2011) on the freedoms of opinion and expression, para. 9, available at www2.ohchr.org/english/bodies/hrc/docs/GC34.pdf; Manfred Nowak, *U.N. Covenant on Civil and Political Rights: CCPR Commentary* (1993).

[22] Nowak, *U.N. Covenant on Civil and Political Rights*.

[23] *Yong Joo-Kang v. Republic of Korea*, Human Rights Committee communication No. 878/1999, 16 July 2003 (CCPR/C/78/D/878/1999).

intelligence may also reinforce the assumption that search results generated on a particular platform are an objective presentation of factual information.

26. The issues that market dominance raises in the field of artificial intelligence-assisted curation therefore test historical understandings of how content curation affects or does not affect the capacity to form an opinion. The novelty of the issues raised, coupled with the general lack of jurisprudence concerning interferences with the right of opinion, provide more questions than answers about the human rights impact of artificial intelligence-assisted curation in the contemporary digital environment. Nevertheless, these questions should drive rights-oriented research into the social, economic and political effects of artificial intelligence-assisted curation. Companies should, at the very least, provide meaningful information about how they develop and implement criteria for curating and personalizing content on their platforms, including policies and processes for detecting social, cultural or political biases in the design and development of relevant artificial intelligence systems.

C. Right to freedom of expression

27. Article 19 (2) of the Covenant guarantees an expansive right to "seek, receive and impart information and ideas of all kinds", one which must be protected and respected regardless of frontiers or type of media. Enjoyment of the right to freedom of expression is intimately related to the exercise of other rights and foundational to the effective functioning of democratic institutions and, accordingly, the protection, respect and promotion of the right to freedom of expression entails the duty to include the promotion of media diversity and independence and the protection of access to information.[24]

28. Unlike the right to form and hold opinions, the rights to express and access information and ideas may be subject to restrictions under limited circumstances (article 19 (3) of the Covenant). Restrictions must meet the standards of legality, meaning that they are publicly provided by a law that meets standards of clarity and precision and are interpreted by independent judicial authorities; necessity and proportionality, meaning that they are the least intrusive measure necessary to achieve the legitimate interest at hand and do not imperil the essence of the right; and legitimacy, meaning that they must be in pursuit of an enumerated legitimate interest, namely, the protection of rights or reputations of others, national security or public order, or public health or morals (A/HRC/38/35, para. 7). Within this framework, expression rights can also be restricted pursuant to article 20 (2) of the Covenant, which requires States to prohibit "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence", but restrictions must still satisfy the cumulative conditions of legality, necessity and legitimacy.[25]

29. The complexity of decision-making inherent in content moderation may be exacerbated by the introduction of automated processes. Unlike humans, algorithms are today not capable of evaluating cultural context, detecting irony or conducting the critical analysis necessary to accurately identify, for example, "extremist" content or

[24] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Organization for Security and Co-operation in Europe Representative on Freedom of the Media, Organization of American States Special Rapporteur on freedom of expression and African Commission on Human and Peoples' Rights Special Rapporteur on freedom of expression and access to information, "Joint Declaration on freedom of expression and 'fake news'", disinformation and propaganda", 3 March 2017. Available at https://www.osce.org/fom/302796; see also Human Rights Committee, general comment No. 34 (2011) on the freedoms of opinion and expression; A/HRC/29/32, para. 61 and A/HRC/32/38, para. 86.

[25] Human Rights Committee, general comment No. 34 (2011) on the freedoms of opinion and expression, para. 50.

hate speech[26] and are thus more likely to default to content blocking and restriction, undermining the rights of individual users to be heard as well as their right to access information without restriction or censorship.

30. In an artificial intelligence-governed system, the dissemination of information and ideas is governed by opaque forces with priorities that may be at odds with an enabling environment for media diversity and independent voices. Relevantly, the Human Rights Committee has found that States should "take appropriate action ... to prevent undue media dominance or concentration by privately controlled media groups in monopolistic situations that may be harmful to a diversity of sources and views".[27]

31. Users also lack access to the rules of the game when it comes to artificial intelligence-driven platforms and websites. A lack of clarity about the extent and scope of artificial intelligence and algorithmic applications online prevent individuals from understanding when and according to what metric information is disseminated, restricted or targeted. Small concessions to addressing this problem such as selective identification of sponsored search results,[28] or social media platforms highlighting when advertising is paid for by political actors, may contribute slightly to helping users to understand the rules of the information environment, but these neither capture nor resolve the concerns around the scale at which algorithmic processes are shaping that environment.

32. Even when individuals are informed about the existence, scope and operation of artificial intelligence systems, those systems may frustrate efforts at transparency and suitability. To date, no sophisticated and scalable means for scrutinizing and making transparent the technical underpinnings of automated decisions in the online sphere have been developed.[29] This means that individuals will often have their expression rights adversely affected without being able to investigate or understand why, how or on what basis.

D. Right to privacy

33. The right to privacy often acts as a gateway to the enjoyment of freedom of opinion and expression.[30] Article 17 of the Covenant protects the individual against "arbitrary or unlawful interference with his or her privacy, family, home or correspondence" and "unlawful attacks on his or her honour and reputation" and provides that "everyone has the right to the protection of the law against such interference or attacks". The Office of the High Commissioner for Human Rights and the Human Rights Council have emphasized that any interference with privacy must meet standards of legality, necessity and proportionality (A/HRC/27/37, para. 23 and Human Rights Council resolution 34/7, para. 2).

34. Artificial intelligence-driven decision-making systems depend on the collection and exploitation of data, ranging from ambient, non-personal data to personally

[26] Council of Europe, *Algorithms and Human Rights*.

[27] Human Rights Committee, general comment No. 34 (2011) on the freedoms of opinion and expression, para. 40.

[28] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York, New York University Press, 2018).

[29] Mike Ananny and Kate Crawford, "Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability", *New Media and Society*, vol. 20, No. 3 (13 December 2016). Available at http://journals.sagepub.com/doi/abs/10.1177/1461444816676645?journalCode=nmsa.

[30] See A/HRC/29/32, para. 16, General Assembly resolution 68/167 and Human Rights Council resolution 20/8.

identifiable information, with the vast majority of data used to feed artificial intelligence systems being somewhere in the middle — data that are inferred or extracted from personal data, or personal data that have been anonymized (often imperfectly). Companies use data derived from online profiling and digital fingerprinting, procure datasets from third parties such as data brokers and derive new data from vast aggregated datasets to feed artificial intelligence systems. Artificial intelligence-driven consumer products and autonomous systems are frequently equipped with sensors that generate and collect vast amounts of data on individuals within their proximity[31] and artificial intelligence methods on social media platforms are used to infer and generate sensitive information about people that they have not provided or confirmed, such as sexual orientation, family relationships, religious views, health conditions or political affiliation.

35. Artificial intelligence challenges traditional notions of consent, purpose and use limitation, transparency and accountability — the pillars upon which international data protection standards rest.[32] Because artificial intelligence systems work by exploiting existing datasets and creating new ones, the ability of individuals to know, understand and exercise control over how their data are used is deprived of practical meaning in the context of artificial intelligence. Once data are repurposed in an artificial intelligence system, they lose their original context, increasing the risk that data about individuals will become inaccurate or out of date and depriving individuals of the ability to rectify or delete the data. Artificial intelligence-based systems are being used to make consequential decisions using those data, some of which profoundly affect people's lives,[33] and yet, individuals have few avenues to exercise control over data that have been derived from their personal data, even as anonymization techniques continue to suffer from inadequacies.

E. Obligation of non-discrimination

36. Non-discrimination is an intrinsic principle of human rights law, existing not only as a qualifier on the obligations of States to ensure enjoyment of all other human rights without discrimination, but also, as enshrined in article 26 of the Covenant, as a stand-alone guarantee of equality before the law and equal protection of the law. States are under a clear obligation to "prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status". Thus, articles 17 and 19 incorporate individual rights to freedom from discrimination in the holding and forming of opinions, the expression of and access to ideas and information, and the exercise of privacy and the protection of personal data.

37. The potential for artificial intelligence to embed and perpetuate bias and discrimination extends to discrimination in the exercise of freedom of opinion and expression. Moderation algorithms may fail to take into account cultural, language or gender-based contexts and sensitivities, or public interest in the content.[34] Artificial intelligence-driven newsfeeds may perpetuate and reinforce discriminatory attitudes,

[31] Article 19 and Privacy International, "Privacy and freedom of expression".

[32] Human Rights Committee, general comment No. 16: Article 17 (1988) on the right to privacy, para. 10.

[33] Article 19 and Privacy International, "Privacy and freedom of expression".

[34] This has led to, for example, the removal of historical photographs with particular cultural significance. See Julia Carrie Wong, "Mark Zuckerberg accused of abusing power after Facebook deletes 'napalm girl' post", *The Guardian*, 9 September 2016. Available at https://www.theguardian.com/technology/2016/sep/08/facebook-mark-zuckerberg-napalm-girl-photo-vietnam-war; see also A/HRC/38/35, para. 29.

while artificial intelligence profiling and advertising systems have demonstrably facilitated discrimination along racial, religious and gender lines.[35] "Autocomplete" artificial intelligence functions have also produced racially discriminatory results.[36]

38. A number of factors ingrain bias into artificial intelligence systems, increasing their discriminatory potential. These include the way in which artificial intelligence systems are designed, decisions as to the origin and scope of the datasets on which these systems are trained, societal and cultural biases that developers may build into those datasets, the artificial intelligence models themselves and the way in which the outputs of the artificial intelligence model are implemented in practice. For example, facial recognition applications suffer from being grounded in predominantly white, male datasets, with errors occurring in up to 20 per cent of the time for women and people with darker skin colours.[37] When such systems are used to, for example, categorize images available through a search engine, their discriminatory potential can translate into concrete interferences with individuals' exercise of their rights to seek, receive and impart information and freely assemble or associate.

F. Right to an effective remedy

39. Human rights law guarantees individuals a remedy determined by competent judicial, administrative or legislative authorities (article 2 (3) of the Covenant). Remedies must be known by and accessible to anyone who has had their rights violated; must involve prompt, thorough and impartial investigation of alleged violations;[38] and must be capable of ending ongoing violations (A/HRC/27/37, paras. 39–41).

40. Artificial intelligence systems often interfere with the right to a remedy. First, individual notice is almost inherently unavailable. In almost all applications of artificial intelligence technology in the information environment, individuals are not aware of the scope, extent or even existence of the algorithmic decision-making processes that may have an impact on their enjoyment of rights to opinion and expression. The second and more challenging aspect is the scrutability of the artificial intelligence system itself. The logic behind an algorithmic decision may not be evident even to an expert trained in the underlying mechanics of the system. Although it is logical to assume that more transparency around artificial intelligence systems would enable greater scrutiny, algorithmic transparency does not necessarily equate with intelligible explanations of decision-making processes. Algorithms may obscure that a consequential decision has been taken or be so complex and context-dependent as to frustrate explanation. The situation is further complicated because companies

[35] Julia Angwin, Madeleine Varner and Ariana Tobin, "Facebook enabled advertisers to reach 'Jew haters'", ProPublica, 14 September 2017. Available at https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters; Ariana Tobin, "Why we had to buy racist, sexist, xenophobic, ableist and otherwise awful Facebook ads," ProPublica, 27 November 2017. Available at https://www.propublica.org/article/why-we-had-to-buy-racist-sexist-xenophobic-ableist-and-otherwise-awful-facebook-ads.

[36] Paris Martineau, "YouTube's search suggests racist autocompletes", The Outline, 13 May 2018. Available at https://theoutline.com/post/4536/youtube-s-search-autofill-suggests-racist-results?zd=1&zi=3ygzt6hw.

[37] Joy Buolamwini, "The dangers of supremely white data and the coded gaze", presented at Wikimania 2018, Cape Town. Available at https://www.youtube.com/watch?v=/SJXKoD6mA8&feature=youtu.be.

[38] Human Rights Committee, general comment No. 31 (2004) on the nature of the general legal obligation imposed on States parties to the Covenant, para. 15.

operating in the information environment frequently update their algorithms;[39] equally, machine-learning applications may change their own rules and algorithms over time.

41. Compounding these concerns is the shift towards the automation of remedy systems themselves, according to which complaints of individual users, either about content moderation decisions or about the adverse human rights impacts of artificial intelligence technologies, are considered and determined by artificial intelligence technologies.[40] Automatic response processes raise concerns about whether complaint redress mechanisms constitute an effective remedy, given the lack of discretion, contextual analysis and independent determination built into such processes.[41]

G. Legislative, regulatory and policy responses to artificial intelligence

42. Many States are now devising national artificial intelligence strategies in order to explore and develop policies and initiatives designed to maximize the potential benefits of artificial intelligence for their citizens.[42] Although no State has yet to propose a comprehensive law or regulation of artificial intelligence, there are reasons to be cautious about such an approach, which may be ill-suited to such an innovative field and may compensate for lack of detail with overly restrictive or overly permissive provisions. Sectoral regulation may be preferable although, arguably, existing law and regulation, for example in the field of data protection, could be flexible and available without the need to legislate further.

43. At the same time, States should ensure that artificial intelligence is developed in keeping with human rights standards. Any efforts to develop State policy or regulation in the field of artificial intelligence should ensure consideration of human rights concerns.[43] The rights to freedom of opinion and expression, in particular, are often excluded from public and political debates on artificial intelligence, which, to the extent that they tackle human rights issues, tend to focus on bias and discrimination in service delivery.

44. Because the development of effective artificial intelligence systems depends on the acquisition of large datasets as well as long-term investment in technological capabilities, private sector entities are likely to dominate in terms of development, production and capacity, leading to increased public sector reliance on companies for access to artificial intelligence systems. This increases the likelihood and public perception that corporate and government interests will become increasingly intertwined. This is especially the case in the information environment, in which governments are often users — of social media platforms, search engines and other technologies — rather than providers. The alignment of public and private interests does not in itself create human rights interferences but raises concerns about transparency and accountability. As private development of artificial intelligence

[39] Barry Schwartz, "Google: we make thousands of updates to search algorithms each year", Search Engine Roundtable, 5 June 2015. Available at https://www.seroundtable.com/google-updates-thousands-20403.html.

[40] Council of Europe, *Algorithms and Human Rights*.

[41] Pei Zhang, Sophie Stalla-Bourdillon and Lester Gilbert, "A content-linking-context model for 'notice-and-take-down' procedures", *WebSci '16*, May 2016. Available at http://takedownproject.org/wp-content/uploads/2016/04/ContentLinkingModelZhangStallaGilbert.pdf.

[42] Tim Dutton, "An overview of national AI strategies", Medium, 28 June 2018. Available at https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

[43] It is concerning, for example, that a parliamentary committee in the United Kingdom of Great Britain and Northern Ireland issued a 200-page report that does not mention human rights even once. See United Kingdom of Great Britain and Northern Ireland House of Lords Select Committee on Artificial Intelligence, "AI in the United Kingdom".

proceeds, there is a very real risk that States will delegate increasingly complex and onerous censorship and surveillance mandates to companies.

45. Any State attempts to articulate law or policy in the field of artificial intelligence should speak both to public and private sector artificial intelligence applications, rather than simply focusing on public sector artificial intelligence regulation. As the Council of Europe concluded, "Issues related to algorithmic governance and/or regulation are public policy prerogatives and should not be left to private actors alone."[44] State approaches may involve enhanced transparency and disclosure obligations on companies and robust data protection legislation that addresses artificial intelligence-related concerns.

46. Public and private sector initiatives designed to explore and integrate ethics into the procurement, design, deployment and implementation of artificial intelligence systems are proliferating. I strongly encourage the integration of human rights concerns into these efforts. The private sector's focus on and the public sector's push for ethics often imply resistance to human rights-based regulation.[45] While ethics provide a critical framework for working through particular challenges in the field of artificial intelligence, it is not a replacement for human rights, to which every State is bound by law. Companies and governments should ensure that human rights considerations and responsibilities are firmly integrated into all aspects of their artificial intelligence operations even as they are developing ethical codes and guidance.[46]

IV. A human rights-based approach to artificial intelligence

47. In recent reports, the mandate holder has set out legal and practical considerations for companies, to put human rights principles at the heart of their content regulation policies, and has detailed both substantive standards and processes that ensure that companies can comply with their human rights responsibilities under the Guiding Principles on Business and Human Rights in all aspects of their operations. That same framework frames the approach offered here with respect to artificial intelligence technologies. The substantive standards and processes proposed below apply to companies, in their capacity as actors that design, deploy and implement artificial intelligence systems, and to States, which have an obligation to refrain from interfering with human rights in their own adoption and use of artificial intelligence systems. These standards and processes are designed to ensure that human rights law is placed at the heart of advancements in the field of artificial intelligence. Two fundamental principles are woven throughout the standards and processes offered: the need to protect and respect individual agency and autonomy, a key precondition to the exercise of the right to freedom of opinion and expression; and the importance of meaningful disclosure on the part of public and sector actors, defined by open and innovative efforts to explain artificial intelligence technologies to the public and facilitate their scrutiny.

---

[44] Council of Europe, *Algorithms and Human Rights*.

[45] Ben Wagner, "Ethics as an escape from regulation: from ethics-washing to ethics-shopping?", in *Being Profiling: Cogitas Ergo Sum*, Mireille Hildebrandt, ed. (Amsterdam University Press (forthcoming)).

[46] Article 19 and Privacy International, "Privacy and freedom of expression".

A. Substantive standards for artificial intelligence systems

48. Companies should orient their standards, rules and system design around universal human rights principles (A/HRC/38/35, paras. 41–43). Public-facing terms and guidelines should be complemented by internal policy commitments to mainstreaming human rights considerations throughout a company's operations, especially in relation to the development and deployment of artificial intelligence and algorithmic systems. Companies should consider how to elaborate professional standards for artificial intelligence engineers, translating human rights responsibilities into guidance for technical design and operation choices. The development of codes of ethics and accompanying institutional structures may be an important complement to, but not a substitute for, commitments to human rights. Codes and guidelines issued by both public and private sector bodies should emphasize that human rights law provides the fundamental rules for the protection of individuals in the context of artificial intelligence, while ethics frameworks may assist in further developing the content and application of human rights in specific circumstances.

49. Companies and governments must be explicit with individuals about which decisions in the information environment are made by automated systems and which are accompanied by human review, as well as the broad elements of the logic used by those systems. Individuals should also be informed when the personal data they provide to a private sector actor (either explicitly or through their use of a service or site) will become part of a dataset used by an artificial intelligence system, to enable them to factor that knowledge into their decision about whether to consent to data collection and which types of data they wish to disclose.[47] Similarly to the public notices required for the use of closed-circuit television cameras, artificial intelligence systems should actively disclose to individuals (through innovative means such as pop-up boxes) in a clear and understandable manner that they are subject or contributing data to an artificial intelligence-driven decision-making process, as well as meaningful information about the logic involved in the process and the significance of the consequences to the individual.

50. Transparency does not stop with the disclosure to individual users about the existence of artificial intelligence technologies in the platforms and online services they use. Companies and governments need to embrace transparency throughout each aspect of the artificial intelligence value chain. Transparency need not be complex to be effective; even simplified explanations of the purpose, policies, inputs and outputs of an artificial intelligence system can contribute to public education and debate.[48] Rather than grapple with the predicament of making convoluted technical processes legible to lay audiences, companies should strive to achieve transparency through the provision of non-technical insights into a system. To that end, the focus should be on educating individual users about an artificial intelligence system's existence, purpose, constitution and impact, rather than about the source code, training data and inputs and outputs.[49]

51. Radical transparency about the impact of an artificial intelligence system in the information environment requires disclosure of, for example, data on how much content is removed by artificial intelligence systems, how often artificial intelligence-suggested content removals are approved by a human moderator, how often content removals are contested and how often challenges to content removals are upheld.

---

[47] Human Rights Committee, general comment No. 16 (1988) on the right to privacy.

[48] Aaron Rieke, Miranda Bogen and David Robinson, "Public scrutiny of automated decisions: early lessons and emerging methods" (Omidyar and Upturn, 2018).

[49] Rieke, Bogen and Robinson, "Public scrutiny of automated decisions".

Aggregate data illustrating trends in content display should be available for users to inspect, alongside case studies that illustrate why certain content will be prioritized over other content. Disclosure about the sources and beneficiaries of political and commercial advertising is a critical element of radical transparency. Public and private sector actors implementing artificial intelligence-driven systems should also be transparent about the limits of the artificial intelligence system, including for example, confidence measures, known failure scenarios and appropriate limitations on use.[50]

52. Tackling the prevalence of discrimination in artificial intelligence systems is an existential challenge for companies and governments; failure to address and resolve the discriminatory elements and impacts will render the technology not only ineffective but dangerous. There is ample thought leadership and resources for companies and governments to draw on in considering how to address bias and discrimination in artificial intelligence systems; broadly speaking, it necessitates isolating and accounting for discrimination at both the input and output levels. This involves, at a minimum, addressing sampling errors (where datasets are non-representative of society), scrubbing datasets to remove discriminatory data and putting in place measures to compensate for data that "contain the imprint of historical and structural patterns of discrimination"[51] and from which artificial intelligence systems are likely to develop discriminatory proxies. Active monitoring of discriminatory outcomes of artificial intelligence systems is also integral to avoiding and mitigating adverse effects on the human rights of individuals.

B. Processes for artificial intelligence systems

53. **Human rights impact assessments**. Embracing radical transparency throughout the artificial intelligence life cycle requires companies and governments to take steps to permit systems to be scrutinized and challenged from conception to implementation. Human rights impact assessments are one tool that can demonstrate a commitment to addressing the human rights implications of artificial intelligence systems and should be performed prior to procurement, development or use and involve both self-assessment and external review. The think tank AI Now has proposed a public agency algorithmic impact assessment that stipulates that governments should undertake an internal review of artificial intelligence systems as well as facilitate external research review processes to test and verify assumptions and conclusions.[52] Companies should also conduct assessments along similar lines.

54. Public sector procurement of artificial intelligence technologies from private vendors must be accompanied by a public consultation to elicit societal views and input on the design and implementation of the artificial intelligence system before it is acquired. Both companies and governments must conduct meaningful and sustained consultations with civil society, human rights groups, relevant local communities and representatives of historically marginalized or underrepresented populations before developing, procuring or using artificial intelligence systems and technologies.

[50] Amnesty International and Access Now, "Toronto declaration: protecting the right to equality and non-discrimination in machine learning systems", art. 27 (d), 2018. Available at https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/.

[51] Iason Gabriel, "The case for fairer algorithms", Medium, 14 March 2018. Available at https://medium.com/@ethics_Society/the-case-for-fairer-algorithms-c008a1212618.

[52] Dillon Reisman and others, "Algorithmic impact assessments: a practical framework for public agency accountability" (AI Now, 2018). Available at https://ainowinstitute.org/aiareport2018.pdf.

55. **Audits**. Facilitating external review of artificial intelligence systems provides a critical guarantee of rigour and independence in transparency. For this reason, ongoing independent audits should supplement pre-procurement human rights impact assessments as an important mechanism of transparency and accountability in artificial intelligence systems. Private sector actors have raised objections to the feasibility of audits in the artificial intelligence space, given the imperative to protect proprietary technology. While these concerns may be well founded, I agree with AI Now that, especially when an artificial intelligence application is being used by a public sector agency, refusal on the part of the vendor to be transparent about the operation of the system would be incompatible with the public body's own accountability obligations.

56. In any event, innovative suggestions for audits of artificial intelligence technology that permit proprietary secrecy abound: zero-knowledge proofs could conceivably be generated by algorithms to demonstrate that they conform to certain properties, obviating the need to scrutinize the underlying algorithm,[53] or algorithms could be disclosed to expert third parties who would hold them in escrow on the condition of confidentiality, permitting public interest scrutiny but not allowing the algorithm to become public.[54] Government regulators from the domains of telecommunications or competition could be permitted access to artificial intelligence systems on a confidential basis, as already occurs, for example, in the regulation of gambling machines in Australia and New Zealand, in which companies must submit their algorithmic systems to regulatory audit review.[55] Academic literature contains other suggestions for innovative forms of artificial intelligence audits.[56]

57. Each of these mechanisms may face challenges in implementation, especially in the information environment, but companies should work towards making audits of artificial intelligence systems feasible. Governments should contribute to the effectiveness of audits by considering policy or legislative interventions that require companies to make artificial intelligence code auditable, guaranteeing the existence of audit trails and thus greater opportunities for transparency to individuals affected.

58. **Individual autonomy**. Artificial intelligence must not invisibly supplant, manipulate or interfere with the ability of individuals to form and hold their opinions or access and express ideas in the information environment. Respecting individual autonomy means, at the very least, ensuring that users have knowledge, choice and control. Pervasive and hidden artificial intelligence applications that obscure the processes of content display, personalization, moderation and profiling and targeting undermine the ability of individuals to exercise the rights to freedom of opinion, expression and privacy. Companies should be mindful of the adverse human rights impacts that flow from artificial intelligence applications that prioritize commercial or political interests over transparency and individual choice.

59. **Notice and consent**. Companies must ensure that users are fully informed about how algorithmic decision-making shapes their use of a platform, site or service. This can be achieved through education campaigns, pop-up boxes, interstitials and other means of signalling when an artificial intelligence system is determining a user's experience of a search engine, news site or social media platform. State-imposed disclosure requirements may be an appropriate means of protecting notice and

[53] Council of Europe, *Algorithms and Human Rights*.

[54] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, Harvard University Press, 2015).

[55] Council of Europe, *Algorithms and Human Rights*.

[56] Christian Sandvig and others, "Auditing algorithms: research methods for detecting discrimination on Internet platforms", paper presented at Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a pre-conference at the 64th annual meeting of the International Communication Association, Seattle, 22 May 2014.

consent. Individuals also have a right to know when their data are being collected by an artificial intelligence application and whether the data will become part of a dataset that will subsequently inform an artificial intelligence application, as well as the conditions on which that data will be used, stored and deleted.

60. **Remedy**. Adverse impacts of artificial intelligence systems on human rights must be remediable and remedied by the companies responsible. The precondition to the establishment of effective remedy processes is ensuring that individuals know that they have been subject to an algorithmic decision (including one that is suggested by an artificial intelligence system and approved by a human interlocutor) and are equipped with information about the logic behind that decision. Beyond that, companies should ensure human review of requests for remedy, in order to provide an appropriate check on the systems and guarantee accountability. Data should be published on the frequency at which remedial mechanisms are triggered for decisions made by artificial intelligence technologies.

V. Conclusions and recommendations

61. In the present report, I have explored existing and potential impacts of artificial intelligence on the rights to freedom of opinion and expression, positing that artificial intelligence is now a critical part of the information environment, posing benefits and risks to individuals' enjoyment of their rights. I have proposed a conceptual framing for thinking about the obligations of States and responsibilities of companies to uphold these rights in the face of expanding technological capabilities and have suggested concrete measures that could be implemented by both governments and companies to ensure that human rights are respected as the power, reach and scope of artificial intelligence technology grows.

Recommendations for States

62. When procuring or deploying artificial intelligence systems or applications, States should ensure that public sector bodies act consistently with human rights principles. This includes, inter alia, conducting public consultations and undertaking human rights impact assessments or public agency algorithmic impact assessments prior to the procurement or deployment of artificial intelligence systems. Particular attention should be given to the disparate impact of such technologies on racial and religious minorities, political opposition and activists. Government deployment of artificial intelligence systems should be subject to regular audits by external, independent experts.

63. States should ensure that human rights are central to private sector design, deployment and implementation of artificial intelligence systems. This includes updating and applying existing regulation, particularly data protection regulation, to the artificial intelligence domain, pursuing regulatory or co-regulatory schemes designed to require businesses to undertake impact assessments and audits of artificial intelligence technologies and ensuring effective external accountability mechanisms.[57] Where applicable, sectoral regulation of particular artificial intelligence applications may be necessary and effective for the protection of human rights. To the extent that such restrictions introduce or facilitate interferences with freedom of expression, States should ensure that they are necessary and proportionate to accomplish a legitimate

[57] Wagner, "Ethics as an escape from regulation".

objective in accordance with article 19 (3) of the Covenant. Artificial intelligence-related regulation should also be developed through extensive public consultation involving engagement with civil society, human rights groups and representatives of marginalized or underrepresented end users.

64. **States should create a policy and legislative environment conducive to a diverse, pluralistic information environment. This includes taking measures to ensure a competitive field in the artificial intelligence domain. Such measures may include the regulation of technology monopolies to prevent the concentration of artificial intelligence expertise and power in the hands of a few dominant companies, regulation designed to increase interoperability of services and technologies, and the adoption of policies supporting network neutrality and device neutrality.**[58]

Recommendations for companies

65. **All efforts to formulate guidelines or codes on ethical implications of artificial intelligence technologies should be grounded in human rights principles. All private and public development and deployment of artificial intelligence should provide opportunities for civil society to comment. Companies should reiterate in corporate policies and technical guidance to engineers, developers, data technicians, data scrubbers, programmers and others involved in the artificial intelligence life cycle that human rights responsibilities guide all of their business operations and that ethical principles can assist by facilitating the application of human rights principles to specific situations of artificial intelligence design, deployment and implementation. In particular, the terms of service of platforms should be based on universal human rights principles.**

66. **Companies should make explicit where and how artificial intelligence technologies and automated techniques are used on their platforms, services and applications. The use of innovative means to signal to individuals when they are subject to an artificial intelligence-driven decision-making process, when artificial intelligence plays a role in displaying or moderating content or when individuals' personal data may be integrated into a dataset that will be used to inform artificial intelligence systems is critical to giving users the notice necessary to understand and address the impact of artificial intelligence systems on their enjoyment of human rights. Companies should also publish data on content removals, including how often removals are contested and challenges to removals are upheld, as well as data on trends in content display, alongside case studies and education on commercial and political profiling.**

67. **Companies must prevent and account for discrimination at both the input and output levels of artificial intelligence systems. This involves ensuring that teams designing and deploying artificial intelligence systems reflect diverse and non-discriminatory attitudes and prioritizing the avoidance of bias and discrimination in the choice of datasets and design of the system, including by addressing sampling errors, scrubbing datasets to remove discriminatory data and putting in place measures to compensate for such data. Active monitoring of discriminatory outcomes of artificial intelligence systems is also essential.**

[58] Autorité de régulation des communications électroniques et des postes, *Devices, the Weak Link in Achieving an Open Internet* (2018). Available at https://www.arcep.fr/uploads/tx_gspublication/rapport-terminaux-fev2018-ENG.pdf.

68. **Human rights impact assessments and public consultations should be carried out during the design and deployment of new artificial intelligence systems, including the deployment of existing systems in new global markets. Public consultations and engagement should occur prior to the finalization or roll-out of a product or service, in order to ensure that they are meaningful, and should encompass engagement with civil society, human rights defenders and representatives of marginalized or underrepresented end users. The results of human rights impact assessments and public consultations should themselves be made public.**

69. **Companies should make all artificial intelligence code fully auditable and should pursue innovative means for enabling external and independent auditing of artificial intelligence systems, separately from regulatory requirements. The results of artificial intelligence audits should themselves be made public.**

70. **Individual users must have access to remedies for the adverse human rights impacts of artificial intelligence systems. Companies should put in place systems of human review and remedy to respond to the complaints of all users and appeals levied at artificial intelligence-driven systems in a timely manner. Data on the frequency at which artificial intelligence systems are subject to complaints and requests for remedies, as well as the types and effectiveness of remedies available, should be published regularly.**